Transformers as Soft Reasoners over Language

Year: 2020 | Citations: 412 | Authors: Peter Clark, Oyvind Tafjord, Kyle Richardson

---

Abstract

Beginning with McCarthy's Advice Taker (1959), AI has pursued the goal of providing a system with explicit, general knowledge and having the system reason over that knowledge. However, expressing the knowledge in a formal (logical or probabilistic) representation has been a major obstacle to this research. This paper investigates a modern approach to this problem where the facts and rules are provided as natural language sentences, thus bypassing a formal representation. We train transformers to reason (or emulate reasoning) over these sentences using synthetically generated data. Our models, that we call RuleTakers, provide the first empirical demonstration that this kind of soft reasoning over language is learnable, can achieve high (99%) accuracy, and generalizes to test data requiring substantially deeper chaining than seen during training (95%+ scores). We also demonstrate that the models transfer well to two hand-authored rulebases, and to rulebases paraphrased into more natural language. These findings are significant as it suggests a new role for transformers, namely as limited "soft theorem provers" operating over explicit theories in language. This in turn suggests new possibilities for explainability, correctability, and counterfactual reasoning in question-answering.